CRTgeeDR: An R Package for Doubly Robust Generalized Estimating Equations Estimations in Cluster Randomized Trials with Missing Data

Mélanie Prague, Rui Wang, Victor de Gruttola

Harvard University
Harvard University Biostatistics Working Paper Series

Year 2016 *Paper* 200

CRTgeeDR: An R Package for Doubly Robust Generalized Estimating Equations Estimations in Cluster Randomized Trials with Missing Data

Melanie Prague* Rui Wang†

Victor De Gruttola‡

*Harvard T.H. Chan School of Public Health, mprague@hsph.harvard.edu

†Harvard T.H. Chan School of Public Health and Brigham & Women's Hospital, rwang@hsph.harvard.edu

‡Harvard T.H. Chan School of Public Health, degrut@hsph.harvard.eduedu

http://biostats.bepress.com/harvardbiostat/paper200

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

CRTgeeDR: an R package for Doubly Robust Generalized Estimating Equations Estimations in Cluster Randomized Trials with missing data.

Melanie Prague
Harvard T.H. Chan
Biostatistics department

Rui Wang
Harvard T.H. Chan
Brigham and Women's Hospital

Victor De Gruttola
Harvard T.H. Chan
Biostatistics department

Abstract

Semi-parametric approaches based on generalized estimating equation (GEE) are widely used to analyse correlated outcomes. Most available softwares had been developed for longitudinal settings. In this paper, we present a R package **CRTgeeDR** for estimating parameters in marginal regression in cluster randomized trials (CRTs). Theory for adjusting for missing at random outcomes by inverse-probability weighting methods (IPW) based on the use of a propensity score had been largely studied and implemented. We exhibit that in CRTs most of the available softwares use an implementation of weights that lead to a bias in estimation if a non-independence working correlation structure is chosen. In **CRTgeeDR**, we solve this problem by using a different implementation while keeping the consistency properties of the IPW. Moreover, in CRTs using an augmented GEE (AUG) allow to improve efficiency by adjusting for treatment-covariate interactions and imbalance in baseline covariates between treatment groups using an outcome model. In **CRTgeeDR**, we extend the abilities of existing packages such as **geepack** and **geeM** to allow such data augmentation. Finally, one may want to combine IPW and AUG in a Doubly Robust (DR) estimator, which lead to consistent estimation when either the propensity score or the outcome model corresponds to the true data generation process (Prague, Wang, Stephens, Tchetgen Tchetgen, and De gruttola 2015). The DR approach is implemented in **CRTgeeDR**. Simulations studies demonstrate the consistency of IPW implemented in **CRTgeeDR** and the gains associated with the use of the DR for analyzing a binary outcome using a logit regression. Finally, we reanalyzed data from a sanitation CRT in developing countries (Guiteras, Levinsohn, and Mobarak 2015a) with the DR approach compared to classical GEE and demonstrated a significant intervention effect.

Keywords: Augmentation, Cluster randomized trial, Correlated data, CRTgeeDR, Doubly Robust, geeM, Generalized Estimating Equation, geepack, inverse probability weighting (IPW), MAR, marginal effect, missing data, R.

1. Introduction

We describe the R (R Core Team 2015) package **CRTgeeDR**, for estimating coefficients of regression in a marginal mean model. The methods is designed to analyze data collected in cluster randomized trials (CRTs) where 1) observations within a cluster may be correlated, 2) observations in separate clusters are independent, 3) a monotone transformation of expectation of the outcome is linearly related to the explanatory variables, 4) the variance is a function of the expectation, and 5) treatment is randomized at a cluster level. The estimation approach generalizes the Generalized Estimating Equation (GEE) (Zeger and Liang 1986) for fitting marginal generalized linear models to clustered data with possibly informative missingness of the outcome. It combines existing methods for accommodating missing data that use inverse probability weighting (IPW) (Robins, Rotnitzky, and Zhao 1995) and for increasing precision of estimation by appropriate use of baseline covariates (AUG) (Stephens, Tchetgen Tchetgen, and DeGruttola 2012). We have developed a method that combines the IPW and the AUG that is doubly robust (DR), which implied that the resulting estimator is consistent if either the outcome model or missing data model are correctly specified – that is, they reflect the true data generation processes Prague *et al.* (2015). Below we illustrate the use of the software on a real dataset and clarify its benefits.

The package **CRTgeeDR** not only implements the DR estimator but also the standard GEE, the IPW and the AUG. Regarding IPW, our package differs from most of those currently available in that it avoids the bias that can result from conventional implementation applied to CRTs. Lin, Rodriguez, and SAS (2015) pointed out that implementation of GEE for complete longitudinal data in the current version of SAS (GENMOD procedure) requires use of an independence correlation structure if the observation of the outcome at one time point depends on covariates obtained at another time point; this problem had been corrected in the new GEE procedure in SAS/STAT 13.2 (SAS Institute Inc. 2015) but not in R to date. Tchetgen Tchetgen, Glymour, Weuve, and Robins (2012) made a similar comment regarding the analysis of incomplete longitudinal data in which time-varying covariates and previous outcome values are needed to model the missingness process. This article clarifies this issue for CRTs and propose an implementation that allows for unbiased IPW (and thus DR) estimation with non-independence working correlation structure.

GEE-based approaches for estimating the coefficients in marginal models, in particular the marginal effect of an intervention, have been implemented in only a limited number of software and R packages for general use. Of note, most of these software have initially been developed to deal with correlated longitudinal data rather than data from CRTs. There are three R packages on CRAN, which will solve GEEs and produce standard errors: whereas **gee** (Carey, Lumley, and Ripley 2012) and **geepack** (Halekoh, Højsgaard, and Yan 2006) are computationally demanding, the package **geeM** allows a fast estimation through the use of sparse matrix representation (McDaniel, Henderson, and Rathouz 2013). When interest lies in adjusting for MAR outcomes using the IPW, all the packages mentioned above require specification of weights. These weights can be computed using packages such as **ipw** (van der Wal and Geskus 2011) or directly plugged from a user-defined function. These approaches require that the missing data process be correctly specified. Some packages, such as **drgee** (Zetterqvist and Sjölander 2015), implement doubly robust approaches for uncorrelated data arising from observational studies. These packages provide estimates that are doubly robust in the sense that the consistency of the parameter estimator from the marginal models is

guaranteed if the model linking the outcome and covariates or the model linking the treatment assignment and covariates correctly reflects the true data generation process. These methods have been extended to deal with missing data with IPW approaches **CausalGAM** (Glynn and Quinn 2010), but these packages are intended for analysis of observational studies, not CRTs. Finally, targeted maximum likelihood estimation (tMLE) method also allows estimation of the marginal additive effect of a treatment. It is implemented in the package **tmle** (van der Laan 2010). A discussion about the differences between GEE-based and tMLE estimation in longitudinal data can be found in Porter, Gruber, van der Laan, and Sekhon (2011). However, the two approaches for CRTs are conceptually different, therefore, the comparison between the two will not be discussed in this article where the focus is on software implementation.

The paper is organized as follows. Section 2 introduces the theory of the Doubly robust estimator and section 3 describes the features of the **CRTgeeDR** and the estimating function denoted `GeedrEstimation`. Section 4 compares the performance of **CRTgeeDR** to **geepack** and **geeM** for the IPW in CRTs and illustrates that the DR is also consistent and more efficient than the IPW. Section 5 illustrates the analysis of a dataset on sanitation in developing countries (Guiteras *et al.* 2015a) and illustrates the benefit of using the DR approach compared to standard GEE. Section 6 presents a discussion.

2. IPW in CRTs and Doubly Robust Estimation

2.1. Notation

Consider a CRTs comprised of n clusters or communities, each with n_i individuals. The cluster sample sizes are assumed fixed and non-informative. Let $\boldsymbol{Y}_i = [Y_{ij}]_{j=1,\ldots,n_i}$ denote the outcome vector for cluster i, some elements of which may be unobserved. Let $R_{ij} = 1$ if Y_{ij} is observed and $R_{ij} = 0$ otherwise. Let $\boldsymbol{X}_{ij} = [X^r_{ij}]_{j=1,\ldots,n_i;r=1,\ldots,P}$ denote the P baseline covariates for subject j in cluster i, which is fully observed. Let A_i be the treatment assigned to cluster i; the indicator for treated condition is $A_i = 1$, and that for control condition is $A_i = 0$. We assume that the probability of treatment assignment is known and fixed to $\pi_a = P(A_i = 1)$. The conditional mean of Y_{ij} is denoted $\mu_{ij} = E(Y_{ij}|X_{ij}, A_i)$, and we let $\boldsymbol{\mu}_i = [\mu_{ij}]_{j=1,\ldots,n_i}$ denote the full vector of means in the i^{th} cluster. We assume the mean structure of Y_{ij} depends on the covariate vector for subject j in cluster i (Robins, Greenland, and Hu 1999), and consider a model for the mean of the form:

$$g(\mu_{ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta}_X + A_i\beta_A,$$

where $g(.)$ is a monotone differentiable link function and $\boldsymbol{\beta} = (\beta_A, \boldsymbol{\beta}_X)$ is a $(P+1)\times 1$ vector of regression coefficients of interest. In this article, we focus on estimation of the marginal effect of an intervention β_A for a binary outcome using the logit link. We assume the variance is $v_{ij} = \text{var}(Y_{ij}|X_{ij}, A_i) = \phi h(\mu_{ij})$, where $h(.)$ is the variance function and ϕ is the dispersion parameter. Thus for our specific example, $v_{ij} = \phi\mu_{ij}(1-\mu_{ij})$ We assume a restricted version of the missing at random (rMAR) assumption, which implies that the missingness indicator R_{ij} is a function only of covariates including treatment condition. Although all the theory would hold for classical MAR assumption, we often need to be more stringent in CRTs because it is difficult to specify the function linking missingness and the observed outcome of other individuals in the same cluster. Thus, the probability of being missing π_{ij}, which we call the

Propensity Score (PS), can be expressed as: $\pi_{ij}(\boldsymbol{X}_{ij}, A_i, \eta_W) = P(R_{ij} = 1|\boldsymbol{X}_{ij}, A_i)$. The parameters η_W are nuisance parameters and must be estimated.

2.2. IPW in CRTs

In presence of rMAR outcome, as in Robins *et al.* (1995), we want to estimate $\boldsymbol{\beta}$ by using Inverse Probability Weighted Generalized Estimating Equation (IPW-GEE). Therefore, we must include a weight matrix $\boldsymbol{W}_i$ to the usual GEE, that is:

$$\boldsymbol{W}_i(\boldsymbol{X}_{ij}, A_i, \eta_W) = \text{diag}\Big(\frac{1}{\pi_{ij}(\boldsymbol{X}_{ij}, A_i, \eta_W)}\Big)_{j=1,\ldots,n_i}.$$

This matrix $\boldsymbol{W}_i(\boldsymbol{X}_{ij}, A_i, \eta_W)$, denoted simply as $\boldsymbol{W}_i$, adjusts the contribution of each individual in a given cluster by upweighting the contribution of individuals who are less likely to be observed. Thus, if the propensity score is correctly specified, i.e. correspond to the true missingness process, the IPW-GEE equation 1 provides consistent estimates:

$$0 = \sum_{i=1}^{n} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{W}_i(\boldsymbol{Y}_i - \boldsymbol{\mu}_i), \tag{1}$$

where $D_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ is a derivative matrix and $\boldsymbol{V}_i$ is the working covariance matrix for the response $\boldsymbol{Y}_i$. In particular, $\boldsymbol{V}_i = \phi\boldsymbol{F}_i^{1/2}\boldsymbol{C}(\boldsymbol{\alpha})\boldsymbol{F}_i^{1/2}$, where $\boldsymbol{F}_i^{1/2} = \text{diag}(h(\mu_{ij}))_{j=1,\ldots,n_i}$ and $\boldsymbol{C}(\boldsymbol{\alpha})$ is the working correlation structure with non-diagonal terms $\boldsymbol{\alpha}$. For example, for an independence correlation structure $\boldsymbol{\alpha}$ is zero; for exchangeable structure, all the elements of $\boldsymbol{\alpha}$ are identical. Parameters $\boldsymbol{\alpha}$ could also depend on the treatment assignment $\boldsymbol{C}(\boldsymbol{\alpha}(A_i))$ but we do not consider this possibility in our implementation. In the package **CRTgeeDR**, we estimate the $\boldsymbol{\alpha}$ and ϕ parameters using moment estimators from the Pearson residuals and the Pearson Chi-Square statistic as in McDaniel *et al.* (2013). In the absence of missing data, $\boldsymbol{W}_i = \boldsymbol{I}$ is set to identity, and the standard GEE is performed by **CRTgeeDR**.

In existing packages such as **geepack** and **geeM**, the Equation 1 is implemented as $0 = \sum_{i=1}^{n} \boldsymbol{D}_i\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$, with $\boldsymbol{V}_i^{-1} = \phi\boldsymbol{F}_i^{1/2}\boldsymbol{W}_i^{1/2}\boldsymbol{C}(\boldsymbol{\alpha})\boldsymbol{W}_i^{1/2}\boldsymbol{F}_i^{1/2}$ to ensure the invertibility of $\boldsymbol{V}_i$. It is easy to verify that when an independence correlation structure is used $\boldsymbol{C}(\boldsymbol{\alpha}) = I$, the two implementations are identical. Therefore, one can always use **geepack** and **geeM** with an independence working correlation structure. On the contrary, if a non-independence working correlation structure is used the IPW consistency do not hold. Let's denote $[\boldsymbol{M}]_{st}$ the non-diagonal term located on the s^{th} row and t^{th} column of the matrix $\boldsymbol{M}$. Then, in the implementation $\boldsymbol{V}_i^{-1} = \phi\boldsymbol{F}_i^{1/2}\boldsymbol{W}_i^{1/2}\boldsymbol{C}(\boldsymbol{\alpha})\boldsymbol{W}_i^{1/2}\boldsymbol{F}_i^{1/2}$, one can write:

$$[\boldsymbol{V}_i^{-1}]_{st} = \phi[\boldsymbol{F}_i^{1/2}]_{ss}[\boldsymbol{F}_i^{1/2}]_{tt}[C(\boldsymbol{\alpha})]_{st}[\boldsymbol{W}_i^{1/2}]_{tt}[\boldsymbol{W}_i^{1/2}]_{ss}.$$

As demonstrated in Prague *et al.* (2015), the proof of consistency for the IPW relies on the fact that, when the PS is correctly modeled, i.e. $\pi_{ij} = P(R_{ij}|X_{ij}, A_i)$:

$$\boldsymbol{M} = E[\phi\boldsymbol{F}_i^{1/2}\boldsymbol{C}(\boldsymbol{\alpha})\boldsymbol{F}_i^{1/2} - \phi\boldsymbol{F}_i^{1/2}\boldsymbol{W}_i^{1/2}\boldsymbol{C}(\boldsymbol{\alpha})\boldsymbol{W}_i^{1/2}\boldsymbol{F}_i^{1/2}|\boldsymbol{X}_{i.}, A_i] = 0.$$

The s^{th} row and t^{th} column of the matrix $\boldsymbol{M}$ can be rewritten as:

$$\begin{aligned} E\left[[\boldsymbol{M}]_{st}|\boldsymbol{X}_{i.}, A_i\right] &= E\left[\underbrace{\phi[\boldsymbol{F}_i^{1/2}]_{ss}[\boldsymbol{F}_i^{1/2}]_{tt}[\boldsymbol{C}(\boldsymbol{\alpha})]_{st}}_{\gamma}\left[1-[\boldsymbol{W}_i^{1/2}]_{tt}[\boldsymbol{W}_i^{1/2}]_{ss}\right]|\boldsymbol{X}_{i.}, A_i\right] \\ &= \gamma E\left[1-[\boldsymbol{W}_i^{1/2}]_{tt}[\boldsymbol{W}_i^{1/2}]_{ss}|\boldsymbol{X}_{i.}, A_i\right] \\ &= \begin{cases} \gamma E\left[\frac{\pi_{is}-R_{is}}{\pi_{is}}|\boldsymbol{X}_{i.}, A_i\right] = 0 & \text{if s=t} \\ \gamma E\left[\frac{\sqrt{\pi_{is}}\sqrt{\pi_{it}}-\sqrt{R_{is}}\sqrt{R_{it}}}{\sqrt{\pi_{is}}\sqrt{\pi_{it}}}|\boldsymbol{X}_{i.}, A_i\right] \neq 0 & \text{otherwise} \end{cases}. \end{aligned}$$

When $s \neq t$, $E\Big[[\boldsymbol{M}]_{st}|\boldsymbol{X}_{i.}, A_i\Big] = 0$ if and only if:

- either $\gamma = 0$, which is always the case when $\boldsymbol{C}(\boldsymbol{\alpha})$ is the independence matrix,
- or $\forall(s,t)[\boldsymbol{W}_i^{1/2}]_{ss} = [\boldsymbol{W}_i^{1/2}]_{tt}$, i.e. $\pi_{is} = \pi_{it}$ which means that the weights are defined at a cluster level and not individual-specific, which is usually not the case in CRTs.

Implementation in **CRTgeeDR**, as in Equation 1, allows the use of non-independence working correlations structure for the IPW while retaining consistency of estimators. Indeed:

$$\begin{aligned} &E[\phi\boldsymbol{F}_i^{1/2}\boldsymbol{C}(\boldsymbol{\alpha})\boldsymbol{F}_i^{1/2} - \phi\boldsymbol{F}_i^{1/2}\boldsymbol{C}(\boldsymbol{\alpha})\boldsymbol{F}_i^{1/2}\boldsymbol{W}_i|\boldsymbol{X}_{i.}, A_i] \\ &\quad = \phi\boldsymbol{F}_i^{1/2}\boldsymbol{C}(\boldsymbol{\alpha})\boldsymbol{F}_i^{1/2}E\Big[\underbrace{I-\boldsymbol{W}_i}_{diag(\frac{\pi_{ij}-R_{ij}}{\pi_{ij}})_{j=1,\ldots,n_i}}|\boldsymbol{X}_{i.}, A_i\Big] = 0. \end{aligned}$$

2.3. Augmentation and Doubly Robust Estimation

Recent advances in methods for analysis of data from CRTs have used augmented GEE to improve efficiency of inferences by incorporating baseline covariates to adjust for imbalance arising by happenstance (Stephens *et al.* 2012); we denote this estimator the AUG. They have also been extended to accommodate missing data using an approach based on the IPW; we denote this estimator the DR, whose properties are described in Prague *et al.* (2015). The DR estimating equation is given by :

$$\begin{aligned} 0 &= \sum_{i=1}^{M}\Bigg[\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{W}_i\left(\boldsymbol{Y}_i - \boldsymbol{B}_i(\boldsymbol{X}_{ij}, A_i, \boldsymbol{\eta}_B)\right) \\ &\qquad + \sum_{a=0,1}\pi_a^a(1-\pi_a)^{1-a}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\Big(\boldsymbol{B}_i(\boldsymbol{X}_{ij}, A_i=a, \boldsymbol{\eta}_B) - \boldsymbol{\mu}_i(\boldsymbol{\beta}, A_i=a)\Big)\Bigg]. \qquad (2) \\ &= \boldsymbol{\Phi}(\boldsymbol{Y}_i, \boldsymbol{R}_i, A_i, \boldsymbol{X}_{ij}, \boldsymbol{\beta}, \boldsymbol{\eta}_W, \boldsymbol{\eta}_B). \end{aligned}$$

Each element of the vector $\boldsymbol{B}_i(\boldsymbol{X}_i, A_i=a, \boldsymbol{\eta}_B) = [B_{ij}(\boldsymbol{X}_i, A_i=a, \boldsymbol{\eta}_B)]_{j=1,\ldots,n_i}$ is an arbitrary function linking Y_{ij} with $\boldsymbol{X}_{ij}$ for each treatment arm. The $\boldsymbol{\eta}_B$ are nuisance parameters. The

estimator in Equation 2 is most efficient if $B_{ij}(\boldsymbol{X}_i, A_i = a, \boldsymbol{\eta}_B) = E(Y_{ij}|\boldsymbol{X}_{ij}, A_i = a)$ (Zhang, Tsiatis, and Davidian 2008). In that case we call $B_{ij}(\boldsymbol{X}_i, A_i = a, \boldsymbol{\eta}_B)$ the outcome model (OM) and say that it is correctly specified. If the OM is not correctly specified, i.e. does not correspond to the true data generation process, the estimation remains consistent but one may have a loss in efficiency. Without missing data, $\boldsymbol{W}_i = \boldsymbol{I}$ is set to identity, and the AUG is performed by **CRTgeeDR**. Figure 1 is a flowchart that describes in which situation each estimator should be used.

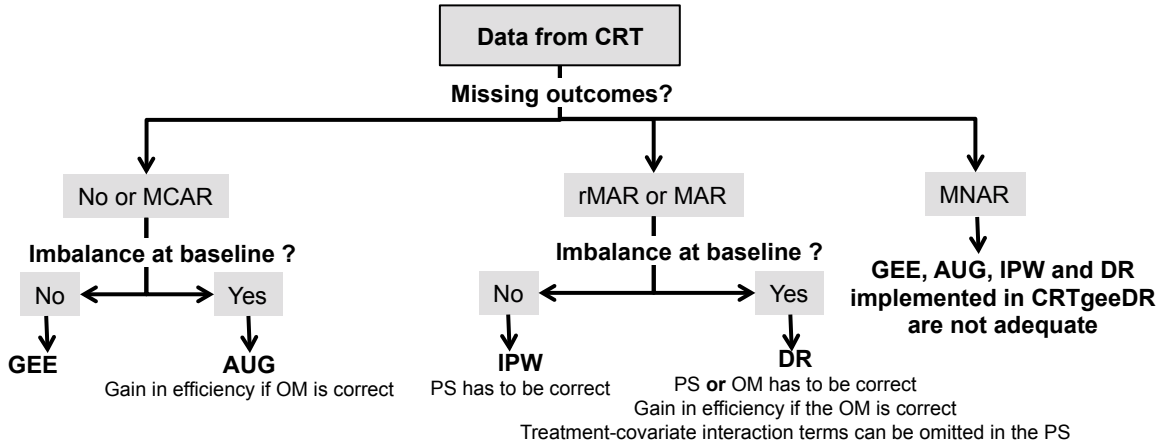

Figure 1: Flowchart describing how to select a consistent and efficient estimator in the package **CRTgeeDR** depending on the presence of missing data (MCAR: missing completely at random, (r)MAR: (restricted) missing at random and MNAR: missing not at random) and imbalance in baseline covariates.

3. The R package CRTgeeDR

The `R` package **CRTgeeDR** contains the functions described below. Examples and description are also available in the package documentation.

- `geeDREstimation` to estimate the regression coefficients in the mean marginal models,
- `getPSPlot` and `getOMPlot`, to plot the distribution of inverse probability weights and the adjustment of outcome model in each treatment group,
- `summary`, to summarize results,
- `fitted` and `predict` to extract fitted or predicted values into a new dataset.

Package **CRTgeeDR** also contains the simulated datasets `data.sim` mimicking the data from a CRT in HIV risk reduction after STI/HIV intervention and documented code to analyse these data.

3.1. The main function for estimation in the package `CRTgeeDR`

The call function for performing estimation is `geeDREstimation`:

```
R> geeDREstimation(formula, id, data = parent.frame(), family = gaussian,
        corstr = "independence", Mv = 1, corr.mat = NULL
        init.beta = NULL, init.alpha = NULL, init.phi = 1, scale.fix = FALSE,
        maxit = 20, tol = 1e-05, print.log = FALSE,
        nametrt = "TRT", nameMISS = "MISSING", nameY = "OUTCOME",
        sandwich = TRUE,sandwich.nuisance = FALSE,
        fay.adjustment = FALSE, fay.bound = 0.75,

        aug = NULL, pi.a = 1/2, model.augmentation.trt = NULL,
        model.augmentation.ctrl = NULL, stepwise.augmentation = FALSE,

        weights = NULL, typeweights = "VW", model.weights = NULL,
        stepwise.weights = FALSE)
```

In this call, the first group of parameters is related to the standard GEE and computation of the variance options, the second group is related to the PS description and IPW options, and finally the third group is related to the OM description and AUG options. The marginal model, to be estimated on the R dataframe `data`, is given in `formula`. The link function, g, depends on the nature of the outcome, which is specified in the attribute `family`. The name of the outcome `nameY`, the clustering variable `id`, the binary treatment `nameTRT` (with the convention 1 is treated and 0 is control), and the missing indicator `nameMISS` must be specified if they differ from default values. The algorithm iterates between the estimation of nuisance parameters and regression parameters with a stopping rule based on stabilization of estimates (tolerance can be set by the user; default is `tol`$= 10^{-5}$ or `max.iter=20`). Depending on the specification or not of the PS and the OM, `geeDREstimation` allows to perform the standard GEE, the IPW, the AUG and the DR approaches. The algorithm is defined as follow:

1. *Determine the PS:* $\pi_{ij}(\boldsymbol{X}_{ij}, A_i, \eta_W) = P(R_{ij}|\boldsymbol{X}_{ij}, A_i)$, π_{ij} for short. Either the π_{ij} are known from prior analysis or by design and the weights can be specified directly in the `weights` attribute. Alternatively one can fit internally a logistic regression of R_{ij} on $(\boldsymbol{X}_{ij}, A_i)$ to compute the PS. In this case, the PS regression formula can be directly entered in `model.weights`. Then a `glm` with logit link function is internally processed with or without variables selection, depending on the value of the `stepwise.weights` attribute. If all of the above are set to NULL or default, no IPW adjustment will be performed. Finally, if despite our cautionary note about the implementation of weights, one wants to use the same implementation as in packages **geepack** and **geeM** or **proc GENMOD** in SAS, then one can set `typeweights="GENMOD"`.

2. *Determine group-specific OM:* $B_{ij}(X_{ij}, A_i = a) = E\left[\boldsymbol{Y}_{ij}|A_i = a, X_{ij}\right]$. Either the $\boldsymbol{B}_i$ are know from prior analysis and can be directly entered in `aug=c(ctrl=`$B_{ij}(X_{ij}, A_i = 0)$`,trt=`$B_{ij}(X_{ij}, A_i = 1)$`)`. Or regression of Y_{ij} on $\boldsymbol{X}_{ij}$ can be fit within each treatment group. In this case, the OM regression formulas can be directly entered in `model.augmentation.trt` and `model.augmentation.ctrl`. Then a `glm` is internally processed with or without variable selection depending on the value of the attribute `stepwise.augmentation`. If all of the above are set to NULL or default, no AUG adjustment will be performed. The probability of treatment assignment which is known in CRTs must be specified in the attribute `pi.a`.

3. *Determine the working correlation structure.* Available structures are `independence`, `exchangeable`, `M-dependent` (using `Mv`), `unstructured`, or `user-defined` (using `corr.mat`). Using the `scale.fix` attribute, the dispersion parameter ϕ can be either estimated or held fixed to a specified value. The implementation of estimation of parameters ϕ and $\boldsymbol{\alpha}$ is the same as in **geeM** and is described in (McDaniel *et al.* 2013).

4. *Obtain initial values.* Either specified by the user (`init.beta`, `init.alpha`, and `init.phi`) or internally defined by fitting a glm under independence to obtain initial value for $\hat{\boldsymbol{\beta}}^{(0)}$ and by setting $\phi^{(0)} = 1$ and $\boldsymbol{\alpha}^{(0)} = 0$.

5. *Enter/Continue the iterative procedure* :

   (a) Use the fit from $\hat{\boldsymbol{\beta}}^{(n)}$ to compute Pearson residuals. Use Pearson residuals based formulas to compute $\phi^{(n+1)}$ the scale parameter (except if `scale.fix`=TRUE) and $\boldsymbol{\alpha}^{(n+1)}$ the parameters in the working correlation matrix.

   (b) Construct the augmented equation given in Equation 2 and solve it numerically using Newton-Raphson algorithm for $\hat{\boldsymbol{\beta}}^{(n+1)}$.

$$\hat{\boldsymbol{\beta}}^{(n+1)} = \hat{\boldsymbol{\beta}}^{(n)} - \left[\frac{\partial \boldsymbol{\Phi}(\boldsymbol{Y}_i, \boldsymbol{R}_i, A_i, \boldsymbol{X}_{ij}, \boldsymbol{\beta}, \boldsymbol{\eta}_W, \boldsymbol{\eta}_B)}{\partial \boldsymbol{\beta}}\right]^{-1}_{\hat{\boldsymbol{\beta}}^{(n)}} \boldsymbol{\Phi}(\boldsymbol{Y}_i, \boldsymbol{R}_i, A_i, \boldsymbol{X}_{ij}, \hat{\boldsymbol{\beta}}^{(n)}, \boldsymbol{\eta}_W, \boldsymbol{\eta}_B)$$

   (c) If $|\hat{\boldsymbol{\beta}}^{(n+1)} - \hat{\boldsymbol{\beta}}^{(n)}| \geq$ `tol` and $n + 1 \leq$ `max.iter` go back to 5 else go to 6.

6. Compute the requested variances of $\hat{\boldsymbol{\beta}}^{(n+1)}$. If, `sandwich` and `sandwich.nuisance` are set to TRUE, classical and nuisance-adjusted (for the estimation of parameters η_W in the PS and η_B in the OM) sandwich estimator of the variance are provided, see Prague *et al.* (2015) for their definition. The nuisance-adjusted version is recommended if the AUG, the IPW or the DR estimator are considered. Finally, a small-sample-adjusted sandwich estimator of the variance can also be computed using Fay's adjustment (Fay and Graubard 2001) setting the attribute `fay.adjustment` to TRUE and specifying the boundary in `fay.bound`.

3.2. Adequacy of the PS and the OM to data

Consistency and efficiency of the DR estimator depend on the correct specification of the PS and the OM, see Prague *et al.* (2015) for theoretical demonstrations. User may want to check the adequacy of the selected OM model to the data by using the function `getOMPlot`, which provide plots to check the `glm` model assumption. The "Residuals vs. Fitted" and the "Scale-location" graphics allow verification of the homogeneity of the variance and the adequacy of the link function. The "Normal Q-Q" checks for the normal distribution of the residuals. The "Residuals vs Leverage" plot allows detection of points that have high leverage on the regression coefficients and that should be investigated as outliers. In the same spirit, the "Cook's distance" and the "Cook's distance vs leverage" provide measures of the effect of deleting a given observation. Of note these graphs are only interpretable for continuous outcome. In addition, for the PS model the function `getPSPlot` provides a histogram of the weights. If weights are too large then the IPW approach is likely to be unstable. Thus,

the user should compute weights externally using, for example, stabilized weights with the associated package **ipw** (van der Wal and Geskus 2011) or other approaches such as described in Wang and Paik (2011). Finally, the user can access the `glm` objects created during the PS and OM steps as outputs named `ps.model`, `om.model.trt` and `om.model.ctrl` from the main function `geeDREstimation`.

4. Simulations

The properties of DR to accommodate complex correlation structure, rMAR outcomes, and presence of imbalance in baseline covariates have already been demonstrated in Prague *et al.* (2015). In this article, we focus on the superiority of implementation of weights in the package **CRTgeeDR** compared to package **geepack** and **geeM**. We focus on a simple example to illustrate that, even in very simple cases, consistency of estimators can be achieved when using exchangeable working correlation structure in existing implementations in R, which use $\boldsymbol{V}_i^{-1} = \phi \boldsymbol{F}_i^{1/2} \boldsymbol{W}_i^{1/2} \boldsymbol{C}(\boldsymbol{\alpha}) \boldsymbol{W}_i^{1/2} \boldsymbol{F}_i^{1/2}$ recall Section 2.2. We simulate data from a cluster randomized trial with 100 communities of 90, 100, or 110 individuals with probability 1/3 for each. The treatment A is randomly assigned with probability $\pi_A = 1/2$. One covariate is of interest: $X_{ij} \sim \mathcal{N}(2, 1)$. We simulate correlated outcome with exchangeable structure, and correlation between individuals is set to 0.05. This is done by using a cluster-level bridge distribution $b_i \sim \mathcal{B}(0.05)$. Data generation process is as follow:

$$\begin{array}{rcl} \text{logit}[P(Y_{ij}|A_i, X_{ij})] & = & -0.5 + 0.3A_i + 0.4X_{ij} + 0.4X_{ij}A_i + b_i, \\ \text{logit}[P(R_{ij} = 1|A_i, X_{ij})] & = & 4.0 - 0.3A_i - 0.8X_{ij} - 0.8X_{ij}A_i \end{array} \tag{3}$$

We simulated R=10,000 replicates. The observed average proportion of missing observations is around 25% and the observed average intraclass correlation is 0.08. The missingness is associated strongly with individual covariates and, thus, the weights differ between individuals in the same cluster. The true value of the odd-ratio for the marginal effect of treatment is computed for each dataset k without missing data by obtaining the counterfactual values with and without treatment under this model:

$$\text{OR}_k = \frac{E(Y_{ij} = 1|A_i = 1)/E(Y_{ij} = 0|A_i = 1)}{E(Y_{ij} = 1|A = 0)/E(Y_{ij} = 0|A_i = 0)}.$$

Then, the true OR is given by $\frac{1}{R}\sum_{k=1}^{R} OR_k$=2.56 with associated parameter for marginal intervention effect in the marginal regression $\beta_A = 0.941$. For each dataset, we first ran the analysis on the dataset without missing data for the standard GEE and the AUG using **CRTgeeDR**. Then, we ran the analysis on the dataset with missing data for the IPW using **geepack** and **geeM** and for the standard GEE, the IPW, the AUG and the DR using **CRTgeeDR**. Two types of DR are presented here, DR1 is the estimator using the correct models for the OM and the PS and Table 1 shows the bias, empirical standard error, sandwich standard error and coverages for each analysis using independence (-I) and exchangeable (-E) working correlation structure. The code to replicate this experiment is available in Web-Supplementary material. The models for the PS and OM for analysis are described in the Table 1. They both can be considered as correctly specified, except for DR2 for which the PS omits treatment-covariate interaction terms and in that sense is misspecified.

The results for standard GEE are unbiased on the datasets without missing data (<0.003 for GEE-I and GEE-E with all packages) and biased in presence of rMAR outcomes (-0.257 for GEE-I and GEE-E with package **CRTgeeDR** and for the other packages as well) implying that

	Independence (-I)				Exchangeable (-E)			
Method	Bias	Emp. SE	SE	Cov.	Bias	Emp. SE	SE	Cov.
No missing data:								
GEE **CRTgeeDR**	0.002	0.102	0.099	94.3	0.002	0.108	0.099	93.2
GEE **geepack**	0.003	0.102	0.101	94.6	0.003	0.102	0.101	94.6
GEE **geeM**	0.002	0.102	0.099	94.3	0.002	0.108	0.099	93.2
AUG **CRTgeeDR**	0.002	0.101	0.099	94.3	0.002	0.109	0.114	95.8
With missing data:								
GEE **CRTgeeDR**	-0.257	0.103	0.177	82.0	-0.256	0.104	0.081	18.1
AUG **CRTgeeDR**	0.249	0.092	0.109	35.7	0.307	0.115	0.139	37.1
IPW **CRTgeeDR**	0.003	0.108	0.106	95.0	0.003	0.118	0.110	93.7
IPW **geepack**	0.008	0.107	0.104	94.8	0.582	0.577	0.357	19.4
IPW **geeM**	0.003	0.108	0.106	95.0	0.098	0.116	0.113	83.5
DR **CRTgeeDR**	0.003	0.107	0.104	94.5	0.004	0.120	0.125	96.1
DR2 **CRTgeeDR**	0.003	0.105	0.102	94.4	0.004	0.118	0.123	96.0

Marginal mean model:
$\text{logit}(\mu_{ij}) = \beta_0 + \beta_A A_i$
PS used for IPW and DR (true):
$\text{logit}(P(R_{ij}|A_i, X_{ij})) = \gamma_0 + \gamma_A A_i + \gamma X_{ij} + \gamma_I X_{ij} A_i$
PS used for DR2 (omitting interactions in PS):
$\text{logit}(P(R_{ij}|A_i, X_{ij})) = \gamma_0 + \gamma_A A_i + \gamma X_{ij}$
OM used for AUG and DR (fitted for each group a):
$\text{logit}(P(Y_{ij}|A_i = a, X_{ij})) = \xi + \xi_A A_i + \xi X_{ij}$

Table 1: Comparison of the standard GEE, the IPW, the AUG and the DR analysis with the package **CRTgeeDR**, **geepack** and **geeM** using independence and exchangeable working correlation structure. True value for the parameter β_A is 0.91 (OR=2.56). The bias, the empirical and the estimated standard errors (SE) and the coverages for parameter $\widehat{\beta_A}$ are computed over 10000 replicates. The true data generation process for outcome and missingness is provided in Equation 3. The PS and OM models for analysis are correctly specified and given in the footnote of the table.

the missingness is informative. Using the IPW-I corrects for this bias regardless of the package used for estimation (0.003 for **CRTgeeDR** and **geeM** and 0.008 for **geepack**). All packages give a similar estimated standard error leading to acceptable coverage close to their nominal value of 95%. When using an exchangeable correlation structure, the coverage (93.7%) remains close to the nominal value for IPW-E using **CRTgeeDR**, but it drops to 19.4% using **geepack** and 83.5% using **geeM**. This is mainly driven by an increase in the bias from 0.003 for **CRTgeeDR** to 0.098 for **geeM** and 0.582 for **geepack** for IPW-E. Using the DR version of **CRTgeeDR** also provides consistent estimates(bias $\leq$ <0.004 for -I and -E). The coverage for the DR is close to or greater than 95% and its use leads most commonly to gains in efficiency. For example, the empirical standard error is 0.108 for IPW-I and 0.107 for DR-I. As demonstrated by DR2, which omits the term $X_{ij}A_i$ in the PS, the doubly robust estimator remains consistent and efficient when the treatment-covariate interactions are not explicitly specified in the PS. As implied by the Prague *et al.* (2015), DR and DR2 have identical properties. But, avoiding the need for treatment-covariate interaction terms in the PS will tend, as demonstrated here,

to have lower standard error leading to more efficiency.

5. Illustration on the sanitation data

In this section, we present a step-by-step analysis of actual CRT data. The data comes from a CRT to investigate the efficacy of alternatives policies on the investment in hygienic latrines in developing countries. A total of 380 communities in rural Bangladesh were assigned to different marketing interventions - community motivation, subsidies, supply side-market, a combination of the three and a control group. Results of this study were published in Science (Guiteras *et al.* 2015a). All the code and data associated with this study are available on dataverse, see url in Guiteras, Levinsohn, and Mobarak (2015b).

	Side-Market supply		Control		All	
Cluster structure						
M	36 (n = 1651)		66 (n = 3186)		100 (n = 4837)	
N_i	49 (15)		48 (16)		48 (16)	
Outcome Y_{ij}	Mean	Missing %	Mean	Missing %	Mean	Missing %
Hygienic Latrine Ownership	34.8%	4.2%	30.3%	3.1%	31.8%	3.5%
Individual-level $\boldsymbol{X}_{ij}^{\text{IND}}$	Mean	Missing %	Mean	Missing %	Mean	Missing %
Report diarrhea	4.3%	0%	4.8%	0%	4.6%	0%
Male	91.1%	<0.01%	90.0%	<0.01%	90.1%	<0.01%
Education	49.2%	0%	45.8%	0%	46.9%	0%
Muslim	83.2%	<0.01%	86.3%	<0.01%	85.2%	<0.01%
Bengali	85.6%	<0.01%	88.5%	<0.01%	87.6%	<0.01%
Agricultor	75.0%	<0.01%	70.2%	<0.01%	71.9%	<0.01%
Stoves	58.2%	<0.01%	62.9%	<0.01%	61.3%	<0.01%
Water Pipes	89.9%	<0.01%	91.3%	<0.01%	90.8%	<0.01%
Phone	64.1%	<0.01%	57.2%	<0.01%	59.5%	<0.01%
Age	39 (13)	<0.01%	39 (14)	<0.01%	39 (14)	<0.01%
Cluster-level $\boldsymbol{X}_{ij}^{\text{C}}$	Mean	Missing %	Mean	Missing %	Mean	Missing %
Village size	230 (120)	0%	270 (190)	0%	256 (170)	0%
Nb doctors	7 (7)	0%	9 (18)	0%	8 (15)	0%
% Landless	41.6 (12)	0%	34.4 (15)	0%	36.9 (15)	0%
% Almost Landless	19.3 (11)	0%	24.0 (8)	0%	22.4 (9)	0%
% Access electricity	59.9 (26)	0%	59.1 (20)	0%	59.4 (22)	0%

Table 2: Description of the Sanitation dataset from (Guiteras *et al.* 2015a) considering only the Side-Market supply and the Control group. Percentages are given for qualitative covariates. Means and standard deviations in parentheses are provided for continuous covariates.

We consider only the comparison of a supply side-market versus control. The published analysis, Guiteras *et al.* (2015a) which used a mixed effect model, showed that the supply side-market alone did not increase the hygienic latrine ownership (+0.3 percentage points, p-value=0.90). Because we want to take advantages of a doubly-robust approach, which is impossible with mixed effect models, we reanalyzed the dataset using GEE approaches. Description of the outcome and variables for adjustment are available in Table 2. Because covariates were missing in less than 0.01% of the observations, we assumed that covariates are MCAR and excluded individuals with missing covariates. The final dataset contains 4774 individuals and 100 clusters. We assume the outcome are rMAR, and conduct the IPW analyses.

As there is some evidence of imbalance in baseline covariates across arms, i.e. the descriptive distributions of covariates in Table 2 are different between treated and control groups, we use the DR approach. We assume that the correlation between any pair of individuals in the same cluster is the same, thus we use an exchangeable working correlation structure. Table 3 presents the PS and OM for analysis, the estimates, the nuisance-adjusted sandwich estimates of the variance, the confidence intervals for the odd-ratios, the p-values, and the computation times for each of these analysis. The PS and OM are fitted using a logistic regression with a linear combination of all the individual-level and cluster-level covariates described in Table 2. Variables for these models are selected using a forward stepwise regression. Inclusion of treatment-covariate interactions in the PS was considered for the IPW but did not affect the results for estimates and is not necessary for DR as demonstrated by (Prague *et al.* 2015). The code for analysis is available in Web-Supplementary material. To illustrate the use of the package **CRTgeeDR**, we provide instructions for the DR estimator:

```
R> DR<-geeDREstimation(OUTCOME~TRT, id=CLUSTER, data = Sanitation,
        family = binomial("logit"), corstr = "exchangeable", typeweights = "VW",
        model.weights = MISSING~TRT+DIARRHEA+...+ELEC_ACCESS,
        model.augmentation.trt = OUTCOME~DIARRHEA+...+ELEC_ACCESS,
        model.augmentation.ctrl = OUTCOME~DIARRHEA+...+ELEC_ACCESS,
        stepwise.weights = TRUE, stepwise.augmentation = TRUE,
        sandwich.nuisance = TRUE)
R> summary(DR)
```

	β_A	Sandwich SE	Nuisance-adjusted SE	$exp(\beta_A)$ OR	IC_{min}	IC_{max}	p-value	time (sec.)
GEE	0.19	0.171	-	1.21	0.87	1.69	0.262	1
IPW	0.19	0.182	0.219	1.21	0.79	1.86	0.386	32
AUG	0.45	0.141	0.176	1.57	1.12	2.22	0.010	11
DR	0.44	0.143	0.183	1.55	1.08	2.21	0.016	20

Marginal mean model: $logit(\mu_{ij}) = \beta_0 + \beta_A A_i$
PS: $\text{logit}(P(R_{ij}|A_i, \boldsymbol{X}_{ij}^{\text{IND}}, \boldsymbol{X}_{ij}^{\text{C}}) = \gamma_0 + \gamma_A A_i + \sum_{k=1}^{10} \gamma_k^{\text{IND}} X_{ijk}^{\text{IND}} + \sum_{k=1}^{5} \gamma_k^{\text{C}} X_{ijk}^{\text{C}}$
OM: $\text{logit}(P(Y_{ij}|A_i = a, \boldsymbol{X}_{ij}^{\text{IND}}, \boldsymbol{X}_{ij}^{\text{C}}) = \xi_0 + \sum_{k=1}^{10} \xi_k^{a\text{IND}} X_{ijk}^{\text{IND}} + \sum_{k=1}^{5} \xi_k^{a\text{C}} X_{ijk}^{\text{C}}$ (for each group a)

Table 3: Effects of the supply side-market vs. control on the probability of hygienic latrine ownership in the sanitation data analysis (Guiteras *et al.* 2015a) using the standard GEE, the IPW adjustment (IPW and DR), and the augmentation for imbalance (AUG and DR) assuming outcomes are rMAR.

For DR the computation time is 20 seconds, which is mainly driven by the computation of the nuisance-adjusted sandwich estimator of the variance (the estimation is < 3 seconds otherwise). Whereas GEE and IPW lead to non-significant effect of supply side-market, augmented approaches do demonstrate effects that are significant at the 0.025 level. Using the DR, we can conclude that there is 55% [8% - 121%] greater change of being a hygienic latrine owner after one year if there is a supply side-market. This effect is significant even using a nuisance-adjusted SE, which is generally larger than the standard sandwich SE due to incorporation of additional variability in estimating nuisance parameters in the PS and

the OM (η_W and η_B). Information about the PS and the OM can be obtained by using the following command lines:

```
R> summary(DR$ps.model)
R> summary(DR$om.model.trt)
R> summary(DR$om.model.ctrl)
R> getPSPlot(DR)
```

	PS π_{ij} (observed)		OM Supply		OM Control	
	Sign	Signif.	Sign	Signif.	Sign	Signif.
Individual-level covariates ($\boldsymbol{X}_{ij}^{\text{IND}}$)						
Supply side-market	-	.				
Report diarrhea	+	**				
Male	+	*	+	.	+	***
Education			+	**	+	***
Muslim					+	***
Bengali			+	***		
Agricultor	+	***				
Stoves	+		+	***	+	***
Water Pipes			+			
Phone	+	***	+	***	+	***
Age	+	***	+	***	+	***
Cluster-level covariates ($\boldsymbol{X}_{ij}^{\text{C}}$)						
Village size			+	.	+	**
Nb doctors	-	***	+	***		
% Landless	-	***	+	***	-	***
% Almost Landless					+	**
% Access electricity					+	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Table 4: Description of covariates from Table 2 selected in the stepwise selection for the PS and the OM (in treated and control) models. Directions of the association and significance of the coefficients are provided.

As noted in Table 3, the estimates for IPW are close to the estimates for GEE. This is because the correction for missingness does not have much impact because only 3.5% of data rare missing, and, as shown in Figure 2, all the non-null weights are close to 1 (mean is 1.035 [1.02; 1.04]). The increased significance of the intervention in the DR analysis compared to GEE is mainly driven by the augmentation. Table 4 displays the covariates among $(\boldsymbol{X}_{ij}^{\text{IND}}, \boldsymbol{X}_{ij}^{\text{C}})$ that are selected by the stepwise procedure for the OM and their significance level. In both groups, households with higher education and economic status (as evidenced by stoves, water pipes, phone, and other factors) are more likely to have a hygienic latrine. For cluster-level covariates the patterns differ more by group; for example, a high number of doctors is positively associated with the hygienic latrine ownership only in the intervention group indicating a potential synergism between the number of doctors and the presence of side-supply markets.

6. Conclusion

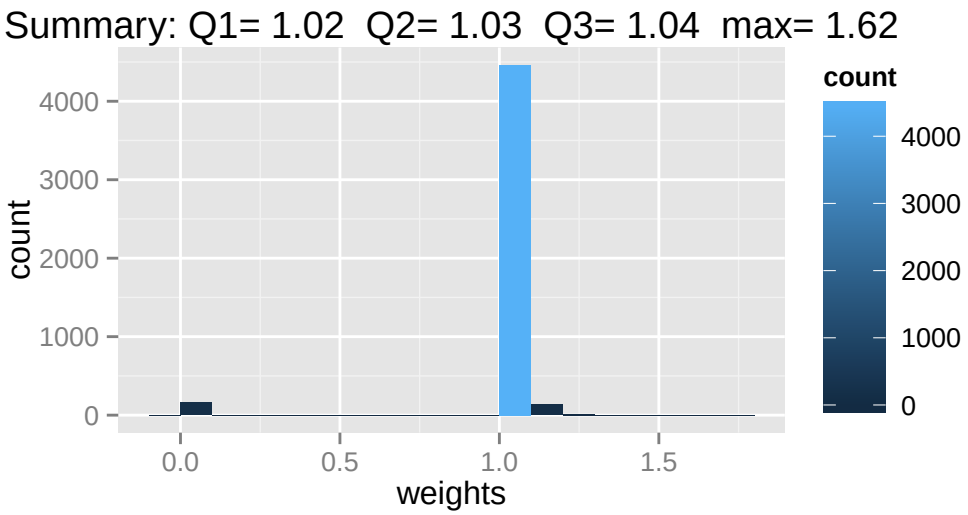

Figure 2: Histogram of weights, which are the diagonal terms of $\boldsymbol{W}_i$, from the regression used for the PS.

We have demonstrated that the IPW can be biased in CRTs if the weights are not implemented as described in Robins *et al.* (1995) and a non-independence working correlation structure is chosen. In particular we raised concerns about the package **geeM** and **geepack** implemented in R. These concerns apply not only for outcome data in CRTs but also to longitudinal outcome data, when the probability that an observations is missing at a given time may depend on time-varying covariates (Tchetgen Tchetgen *et al.* 2012) measured at other times. The **CRTgeeDR** package protects against this bias and allows for adjustment in imbalance in baseline covariates in CRTs. The package can accommodate for a wide range of outcome types, link functions, and working correlation structures. The **CRTgeeDR** package is easy to use and does not require extensive programming. It therefore makes the augmented GEE (AUG) and the Doubly robust (DR) methodology for CRTs (Prague *et al.* 2015) more accessible to applied researchers. Compared to existing packages for GEE approaches, users only need specify two more attributes: a model for the propensity score linking the missingness indicator (R_{ij}) with baseline covariates ($\boldsymbol{X}_{ij}$) and the treatment (A_i), and a model for the outcome linking the outcome (Y_{ij}) with baseline covariates ($\boldsymbol{X}_{ij}$). In presence of rMAR outcomes, if one of these two models is correct regarding the true data generation process, our estimator implemented in **CRTgeeDR** is consistent (unlike GEE and AUG) and generally more efficient than IPW. If the true data generation process linking the missingness (R_{ij}) and the observed outcomes of other individuals in the same cluster ($Y_{ij'}$, with $j' \neq j$) is known, the result above is also valid for MAR outcomes. Finally, although the **CRTgeeDR** package had been designed for CRTs, it can also be used for analysis of correlated longitudinal data from a randomized trial.

Acknowledgement

We thank R. Guiteras for sharing the Sanitation study on the dataverse website. This work

was founded by NIH grants R37 AI 51164 and R01 MH100974. Portions of this research were conducted on the Cluster at Harvard Medical (NIH grant NCRR 1S10RR028832-01).

References

Carey VJ, Lumley T, Ripley B (2012). "gee: Generalized Estimation Equation Solver." *Manual*, pp. 4–13. URL http://CRAN.R-project.org/package=gee,Rpackageversion.

Fay MP, Graubard BI (2001). "Small-Sample Adjustments for Wald-Type Tests Using Sandwich Estimators." *Biometrics*, **57**(4), 1198–1206.

Glynn AN, Quinn KM (2010). "An introduction to the augmented inverse propensity weighted estimator." *Political Analysis*, **18**(1), 36–56.

Guiteras R, Levinsohn J, Mobarak AM (2015a). "Encouraging sanitation investment in the developing world: A cluster-randomized trial." *Science*, **348**(6237), 903–906.

Guiteras R, Levinsohn J, Mobarak M (2015b). "Encouraging Sanitation Investment in the Developing World: A Cluster-Randomized Trial." *Harvard Dataverse;online data*. doi:doi/10.7910/DVN/GJDUTV. URL http://dx.doi.org/10.7910/DVN/GJDUTV.

Halekoh U, Højsgaard S, Yan J (2006). "The R package geepack for generalized estimating equations." *Journal of Statistical Software*, **15**(2), 1–11.

Lin G, Rodriguez R, SAS II (2015). "Weighted Methods for Analyzing Missing Data with the GEE Procedure." *Proceedings of SAS Global Forum*, **Washington DC**(2014 March 23th-26th), paper 166.

McDaniel LS, Henderson NC, Rathouz PJ (2013). "Fast Pure R Implementation of GEE: Application of the Matrix Package." *The R journal*, **5**(1), 181.

Porter KE, Gruber S, van der Laan MJ, Sekhon JS (2011). "The relative performance of targeted maximum likelihood estimators." *The international journal of biostatistics*, **7**(1), 1–34.

Prague M, Wang R, Stephens A, Tchetgen Tchetgen E, De gruttola V (2015). "Accounting for interactions and complex inter-subject dependency for estimating treatment effect in cluster randomized trials with missing at random outcomes." *arXiv preprint arXiv:1507.01822*.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Robins JM, Greenland S, Hu FC (1999). "Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome." *Journal of the American Statistical Association*, **94**(447), 687–700.

Robins JM, Rotnitzky A, Zhao LP (1995). "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data." *Journal of the American Statistical Association*, **90**(429), 106–121.

SAS Institute Inc (2015). *SAS/STAT Software, Version 13.2*. Cary, NC. URL http://www.sas.com/.

Stephens AJ, Tchetgen Tchetgen EJ, DeGruttola V (2012). "Augmented generalized estimating equations for improving efficiency and validity of estimation in cluster randomized trials by leveraging cluster-level and individual-level covariates." *Statistics in medicine*, **31**(10), 915–930.

Tchetgen Tchetgen EJ, Glymour MM, Weuve J, Robins J (2012). "A cautionary note on specification of the correlation structure in inverse-probability-weighted estimation for repeated measures." *Epidemiology*, **23**(4), 644–646.

van der Laan MJ (2010). "Targeted maximum likelihood based causal inference: Part I." *The International Journal of Biostatistics*, **6**(2).

van der Wal WM, Geskus RB (2011). "IPW: an R package for inverse probability weighting." *Journal of Statistical Software*, **43**(13), 1–23.

Wang C, Paik MC (2011). "A Weighting Approach for GEE Analysis with Missing Data." *Communications in Statistics-Theory and Methods*, **40**(13), 2397–2411.

Zeger SL, Liang KY (1986). "Longitudinal data analysis for discrete and continuous outcomes." *Biometrics*, pp. 121–130.

Zetterqvist J, Sjölander A (2015). "drgee: doubly robust generalized estimating equations." *Manual version 1.1.3*. URL http://CRAN.R-project.org/package=drgee.

Zhang M, Tsiatis AA, Davidian M (2008). "Improving efficiency of inferences in randomized clinical trials using auxiliary covariates." *Biometrics*, **64**(3), 707–715.

Affiliation:

Melanie Prague
Department of biostatistics
Harvard T.H. Chan School of Public Health
655 Huntington Ave
Boston, MA 02115
E-mail: mprague@hsph.harvard.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd